Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 1528 – 1532

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Job Opportunity Finding by Text Classification

Shilin Zhang*, Heping Li, Shuwu Zhang

North China University of Technology , Beijing and 100141, China

Abstract

We present a framework to segment words, generate word vectors, train the corpus and make prediction. Based on the text classification technology, we successfully help the disabled persons to acquire job opportunities efficiently in real word. The results show that using this method to build the classifier yields better results than traditional methods. We also experimentally show that careful selection of a subset of features to represent the documents can improve the performance of the classifiers.

Keywords: word segmentation, tfidf, word vector, SVM

1. Introduction

In last years, we have seen an exponential growth in the volume of text documents available on the Web. These Web documents contain rich textual information, but they are so numerous that users find it difficult to obtain useful information from them. This has led to a great deal of interest in developing efficient approaches to organizing these huge resources and assist users in searching the Web. Automatic text classification, which is the task of assigning natural language texts to predefined categories based on their content, is an important research field that can help both in organizing and in finding information in these abundant resources.

Text classification presents many unique challenges and difficulties due to the large number of training cases and features present in the data set. This has led to the development of a number of text classification algorithms, which address these challenges to different degrees. These algorithms include k-NN [1], Naïve Bayes [2], decision tree [3], neural network [4], SVM [5], and Linear Least Squares Fit [6].

* * Corresponding author. Tel.: +86-013521984380

E-mail address: zhangshilin@126.com.

doi:10.1016/j.proeng.2012.01.167

In this paper, we aim to achieve an efficient system to help disabled persons to find job opportunities.

2. Related Work

Text classification is an active research area in information retrieval and machine learning. And several text categorizations have recently been proposed. Furthermore, a feature selection using a hybrid case-based architecture has been proposed by Gentili et al [7] for text categorization where two multi-layer perceptions are integrated into a case-based reasoned. Wermeter has used the document title as the vectors to be used for document categorization [5]. Ruiz and Srinivasan and Calvo and Ceccatto have used the X2 measure to select the relevant features before classifying the text documents using the neural network.

In the scheme, each web page is represented by the term frequency-weighting scheme in the page-preprocessing module and the feature-weighting module. As the dimensionality of a feature vector in the collection set is big, the PCA has been used to reduce it into a small number of principal components in the feature-selecting module.

3. Methodology

In this section, we aim to classify the job information by district and by job type respectively to help the disabled persons easily find their interesting jobs.

3.1. Classifying Job Information By District

Word segmentation and part-of-speech (POS) tagging are important tasks in computer processing of Chinese and other Asian languages. Several models were introduced for these problems, for example, the Hidden Markov Model (HMM) (Rabiner, 1989), Maximum Entropy Model (ME) (Ratnaparkhi and Adwait, 1996).

We adopt a cascaded linear model inspired by the log-linear model (Och and Ney, 2004) widely used in statistical machine translation to incorporate different kinds of knowledge sources. Shown in Fig 1, the cascaded model has a two-layer architecture, with a character based perceptron as the core combined with other real-valued features such as language models.

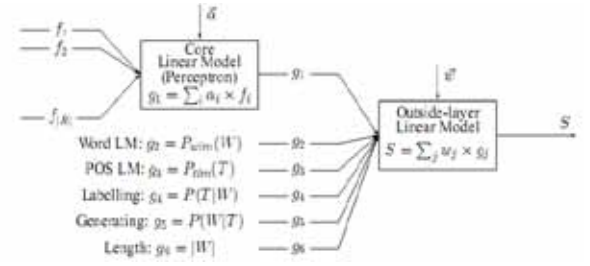

Fig. 1. Structure of Cascaded Linear Model

After segmentation, we get a word vector including the place name to represent a document. Then, we use the vector to classify the document it represents to a pre-defined class. We adopt Native Bayes method to achieve a non-supervised classification. Firstly, we use category names of predefined texts as

class labels. Every category contains the all place names of the category. Then all the class texts can be used as training set, but we avoid the training procedure.

The Fig 2 demonstrates the “JingJinJi” district category information. Like this, we divide all Chinese districts into 14 categories. And we will use it as training sets to classify the preprocessed word vector to the one among the 14 classes.

Fig. 2. category example of jingjinji district

A naive Bayes classifier is a probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers. Abstractly, the probability model for a classifier is a conditional model.

$$p(C \mid F_1 F_2 \cdots \cdots F_n) \tag{1}$$

Here, F1-Fn represent a word vector (a place name string), and the c represent one predefined class name. Over a independent class variable C with a small number of outcomes or classes, conditional on several feature variables F1 through Fn. The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, we write

$$p(C \mid F_1 F_2 \cdots\cdots F_n) = \frac{p(C)p(F_1 F_2 \cdots\cdots F_n \mid C)}{p(F_1 F_2 \cdots\cdots F_n)} \tag{2}$$

We are only interested in the numerator of that fraction, since the denominator does not depend on C and the values of the features Fi are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model. Which can be rewritten as follows, using repeated applications of the definition of conditional probability? Now the "naive" conditional independence assumptions come into play: assume that each feature Fi is conditionally independent of every other feature Fj for j≠i This means that

$$p(F_i \mid C, F_j) = p(F_i \mid C) \tag{3}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C can be expressed like this:

$$p(C \mid F_1 F_2 \cdots\cdots F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i \mid C) \tag{4}$$

As for every document, namely a word vector including a series of place names, we compute the maximum and then we classify the document to the maximum probability class.

3.2. Classification By Job Types

In order to classify the document by job types, we use a two phrase algorithm. In the first phase, we predefined classes of all known job names as class training set. As we can not extract job type name from documents, so we use regular expression to match the document's words and predefined work type names. By this method, we can successfully classify the documents in most scenes. But how can we know all work type names? If a document including a new job type name, the above method will fail. So we should use the following phrase.

4. Experiment Results

To test the proposed system, we collected a data set of job advertisement obtained from the http://www.cdpf.org.cn/ which is official web site of Chinese disabled person's federation, including 5,732 web pages. The types of jobs in the data set are workers(718 documents), designers(116 documents), programmers(953 documents), doctors (1257 documents), accounts (521 documents), managers (126 documents), and the others (962 documents),Among the data set, 4500 documents (about 80%) selected randomly from different classes were used for training data, and the remaining 1232 documents (about 20%) for test data. All the documents are coming from Chinese 31 provinces and cities.

Table 1. Comparision of methods

Method	Precision	Recall
Traditional NB	87%	88%
Traditional SVM	90%	89%
Our Algorithm	91%	92%

Two methods of measuring effectiveness that are widely used in the information extraction research community have been selected to evaluate the metadata extraction including the user preference extraction performance (see Table 1). The methods are: *Precision*: The percentages of actual answers given that are correct. *Recall*: The percentage of possible answers that are correctly extracted.

5. Conclusion

This paper presents a method of automatically classifying Chinese job information into several predefined classes by using text mining techniques for Chinese disabled persons. Based on former researches and the feature of job information, this paper makes some major improvement as follows:

(1) In order to help the Chinese disabled persons to acquire valuable information, we classify the large numbers of job advertisements by district and by job types. And we use different improved algorithms to accomplish it. We did not use the traditional text classification methods. The result shows that our method beats the traditional methods in speed and efficiency.

(2) According to the feature of place names in job advertisements, we propose that documents should be classified respectively by two different procedures: place name extraction and no supervised Native Bayes classifier with the predefined place name sets as training sets in order to improve the accuracy of classification.

Acknowledgements

The work was supported by the Projects in the National Science & Technology Pillar Program with Grant No. 2011BAH16B01.

References

[1] Y. Yang. Expert network. *Effective and efficient learning from human decisions in text categorization and retrieval.* In 17th Annual International ACM SIGIR Conference on Re-search and Development in Information Retrieval (SIGIR'94), pages 13-22, 1994.

[2] A. McCallum and K. Nigam. *A comparison of event models for naïve bayes text classification.* In AAA-98 Workshop on Learning for Text Categorization, 1998.

[3] C. Apte, F. Damerau, and S. Weiss. *Text mining with decision rules and decision trees.* Inproceedings of Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web, 1998.

[4] H.T. Ng, W.B. Goh, and K.L. Low. *Feature selection, perceptron learning, and a usability case study for text categorization.* In 20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), pages 67-73, 1997.

[5] S.Dumais, J. Platt, D. Heckerman, and M. Sahami. *Inductive learning algorithms and representations for text categorization.* In Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management, pages 148-155, 1998.

[6] Y.Yang and C.G. Chute. *An example-based mapping method for text categorization and retrieval.* ACM Transaction on Information Systems (TOIS), 12(3): 252-277, 1994.

[7] G.L. Gentili, M. Marinilli, A. Micarelli, F. Sciarrone. *Text categorization in an intelligent agent for filtering information on the Web.* International Journal of Pattern Recognition and Aritificial Intelligence 15 (3) (2002) 527-549.